Free-Viewpoint Facial Re-Enactment from a Casual Capture

Srinivas Rao
Fyusion Inc.

Rodrigo Ortiz-Cayon
Fyusion Inc.

Matteo Munaro
Fyusion Inc.

Aidas Liaudanskas
Fyusion Inc.

Krunal Chande
Fyusion Inc.

Tobias Bertel
University of Bath

Christian Richardt
University of Bath

Alexander J. B. Trevor*
Robust.AI

Stefan Holzer
Fyusion Inc.

Abhishek Kar*
Google

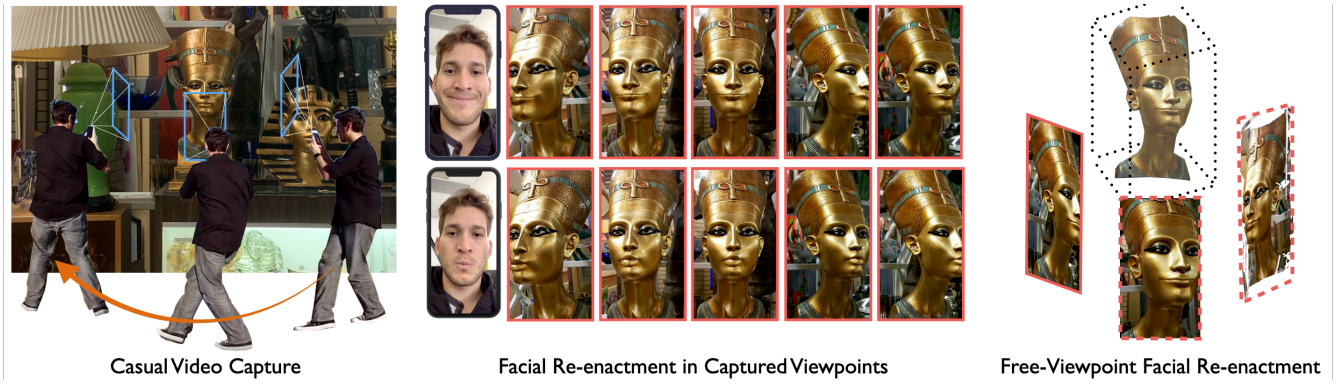

Figure 1: We capture a video around a target subject (the Egyptian bust) and we re-enact the target's face in novel viewpoints. Our re-enactment is driven by an expression sequence of a source subject captured using a custom app running on an iPhone.

ABSTRACT

We propose a system for free-viewpoint facial re-enactment from a casual video capture of a target subject. Our system can render and re-enact the subject consistently in all the captured views. Furthermore, our system also enables interactive free-viewpoint facial re-enactment of the target from novel views. The re-enactment of the target subject is driven by an expression sequence of a source subject, which is captured using a custom app running on an iPhone X. Our system handles large pose variations in the target subject while keeping the re-enactment consistent. We demonstrate the efficacy of our system by showing various applications.

CCS CONCEPTS

• **Computing methodologies** → **Rendering**; **Animation**; **Machine learning**.

KEYWORDS

Facial re-enactment, neural network, image-based rendering

SA '20 Posters, December 04-13, 2020, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8113-0/20/11.
https://doi.org/10.1145/3415264.3425453

ACM Reference Format:
Srinivas Rao, Rodrigo Ortiz-Cayon, Matteo Munaro, Aidas Liaudanskas, Krunal Chande, Tobias Bertel, Christian Richardt, Alexander J. B. Trevor*, Stefan Holzer, and Abhishek Kar*. 2020. Free-Viewpoint Facial Re-Enactment from a Casual Capture. In *SIGGRAPH Asia 2020 (SA '20 Posters), December 04-13, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3415264.3425453

1 INTRODUCTION

Creating life-like photo-realistic avatars is a challenging problem in computer graphics and animation. Facial re-enactment represents a subset of this problem. Previous work often focuses on the case of re-enacting the frontal pose of a human face with RGB or RGB-D input, and many of these require a lengthy capture process for the target subject. More recently, GAN-based methods, like work by Zakharov et al. [2019], show exceptional quality in synthesizing unseen details and poses but they have trouble preserving the identity of the target. Our work is an attempt to navigate this trade-off between identity preservation across all poses and casual capture of the target. Additionally, we attempt to disentangle the desired viewpoint and expression space.

2 SYSTEM OVERVIEW

The goal of our system is to perform free-viewpoint facial re-enactment, i.e. given a source expression sequence, we intend to re-enact the target's face in not only the captured views but also in novel viewpoints. Our system consists of five steps (Figure 2).

*This work was done while the author was working at Fyusion Inc.

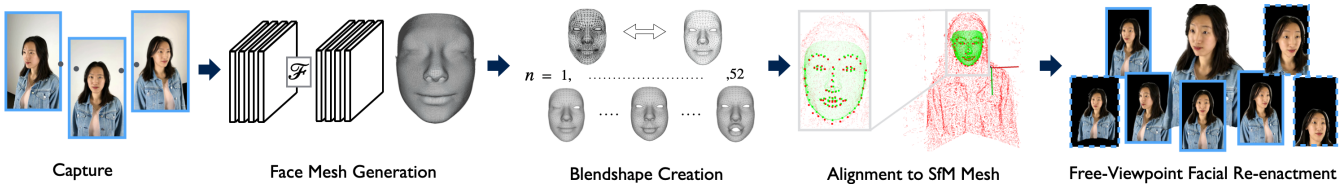

Figure 2: Overview: 1) capture the target subject, 2) generate a face mesh of the target subject, 3) create blendshapes corresponding to target, 4) align face mesh to the SfM mesh, 5) free-viewpoint target re-enactment driven by an expression sequence.

The process starts with a casually captured video of the target with neutral expression. Given multi-view RGB video as input, a server with an NVIDIA GPU processes the following three steps: face mesh generation, blendshape creation, and face mesh alignment. Our final re-enactment is rendered using an OpenGL application running on a MacBook Pro.

2.1 Capture

A fast and simple capture process is key to our system. Our casual capture involves capturing a video of the target with a neutral expression. This typically involves walking around the target subject with any RGB camera, and only takes a few seconds. We also run an optional stabilization and frame selection step to accommodate for the jitter imposed by hand-held capture.

2.2 Face Mesh Generation

After capturing the input video of the target subject, we generate a face mesh corresponding to the target's face. We extended the existing work of Feng et al. [2018] on CNNs for face mesh generation to predict face meshes in frontal pose and a canonical coordinate system. To predict one consistent mesh in our multi-view setup, we combine information from multiple input views using a view-pooling strategy. Our view-pooling step max-pools encoder features produced by all the views, and feeds them into the decoder. We trained this network on a dataset obtained by transforming all face meshes in 300W-LP [Zhu et al. 2016] into a canonical coordinate system such that they are centered at $(128, 128)$ and lie within a $[0, 255]^3$ bounding box with the face pointing along the $+z$ axis. While training, we randomly sample $k \in [1, 8]$ views of a given face from the dataset and employ the view pooling strategy to combine the face mesh predictions.

2.3 Blendshape Creation

We use blendshapes to drive our face re-enactment. To obtain the blendshapes corresponding to the face mesh, we apply a non-rigid deformation to a neutral template blendshape mesh, so it matches the shape of the predicted mesh. We use an off-the-shelf implementation for non-rigid alignment, constrained by matching face landmarks between both meshes. Our template blendshapes consist of 52 expression meshes which is compatible with the 52 blendshape coefficients generated by the ARKit face tracking API. Other expression blendshapes are obtained by transferring the deformation applied to the neutral blendshape. We use the deformation transfer algorithm proposed by Sumner and Popović [2004].

2.4 Alignment to SfM Mesh

The blendshapes generated in the previous step are in the canonical coordinate system. To compute a transformation from the canonical coordinate system to all the captured views, we run a structure-from-motion (SfM) algorithm on the captured video to obtain a static mesh representation of the scene along with camera poses. We align the facial landmarks on the face mesh to the facial landmarks on the SfM mesh and compute a transformation matrix which allows us to establish a central registration between the face mesh, blendshapes, SfM mesh representation, and the captured views. The facial landmarks on the SfM mesh are computed by back-projecting 2D facial landmarks from captured images.

2.5 Free-Viewpoint Facial Re-enactment

We drive all the facial re-enactments using the 52 blendshape coefficients per frame captured by our app running on an iPhone X. Our system enables high-quality free-viewpoint facial re-enactment using image-based rendering. For rendering arbitrary viewpoints, we use the expression mesh corresponding to a given blendshape coefficient as a proxy geometry for the face, and we use a SfM mesh representation for the scene. Optionally, we segment the subject using an off-the-shelf segmentation network to drive realism. Given the proxy geometry, we use our lumigraph implementation [Buehler et al. 2001] for looking up colors from the input viewpoints. This results in a dynamic texture mapping that provides realism to the output as it accommodates for illumination changes across different viewpoints. For synthesizing some missing details, we explore the latent space of the StyleGAN encoder [Karras et al. 2019] for the image corresponding to frontal pose of the target subject. We paste this detail onto a plane, localize the plane onto our SfM mesh representation using facial landmarks, and then project the plane on all captured viewpoints before finally blending it. Please see our supplementary video for example results of our interactive free-viewpoint facial re-enactment approach.

REFERENCES

Chris Buehler, Michael Bosse, Leonard Mcmillan, Steven Gortler, and Michael Cohen. 2001. Unstructured lumigraph rendering. In *SIGGRAPH*.

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *ECCV*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Robert W. Sumner and Jovan Popović. 2004. Deformation Transfer for Triangle Meshes. *ACM Trans. Graph.* 23, 3 (2004), 399–405.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *ICCV*.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Li. 2016. Face Alignment Across Large Poses: A 3D Solution. In *CVPR*.